面向跨语言文本分类与标签推荐的带标签双语主题模型的研究 *

田明杰，崔荣一†

(延边大学 计算机科学与技术学科 智能信息处理研究室，吉林 延吉 133002)

摘　要：针对日渐丰富的跨语言的文字信息资源与新闻报道及科技文献中的多标签数据，为了挖掘跨语言间的相关性及数据属性间的关联性，提出了带标签双语主题模型，应用于跨语言文本分类与标签的推荐。首先，假设科技文献中的关键词与摘要部分有着内容上的相关性，对关键词进行提取，并进行标签化，进而把标签对应于主题模型中的主题，实例化“潜在”的主题；其次，利用带标签双语主题模型对摘要部分进行了训练迭代；最后，对新加入的文档进行跨语言文本分类及标签的推荐。实验结果表明，跨语言文本分类任务中 Micro-F1 达到 94.81%，推荐的标签也较好地体现出语义上的相关性。

关键词：主题模型；标签；跨语言文本分类；标签推荐；潜在主题

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.04.0216

Research on labeled bilingual topic model for cross-lingual text classification and label recommendation

Tian Mingjie, Cui Rongyi†

(*Intelligent Information Processing Laboratory, Dept. of Computer Science & Technology, Yanbian University, Yanji Jilin* 133002, *China*)

Abstract: Aiming at the increasingly rich multi language information resources and multi-label data in news reports and scientific literatures, in order to mining the relevance between languages and the correlation between data, this paper proposed labeled bilingual topic model, applied on cross-lingual text classification and label recommendation. First of all, it could assume that the keywords in the scientific literature are relevant to the **Abstract:** in same article, then extracted the keywords and regarded it as labels, and aligned the labels with topics in topic model, instantiated the “latent” topic. Secondly, trained the **Abstract:** in article through the topic model proposed by this paper. Finally, classified the new documents by cross-lingual text classifier, also recommended the labels. The experiment result show that Micro-F1 measure reaches 94.81% in cross-lingual text classification task, and the recommended labels also reflects the sematic relevance with documents.

Key words: topic model; label; cross-lingual text classification; label recommendation; latent topic

0　引言

随着互联网普及，当今社会进入了信息爆炸的时代。对海量信息进行有效的管理、挖掘利用，有着重要的意义和价值。如今的信息资源，不仅在规模上迅猛增长，资源类型及所使用的语言种类也越来越多样化。语言种类的多样性丰富了信息资源，但是语言间的差异性，不可避免地给用户利用信息资源带来了阻碍。在此背景下，需要跨语言文本分类技术有效地组织多语言信息资源，解决信息杂乱无章的问题。同时，当前众多的新闻报道含有多个标签（label），大多数科技文献与多个关键词（keyword）所链接，其体现的是信息与信息间的相关性。一件事物有不同的属性，同样一个属性也能被标注于不同事物上，这使我们有可能把不同的事物联系起来，挖掘出更多事物间的相关性、更多属性间的关联性。

跨语言文本分类（cross-lingual text classification）是利用已标注好类别的一种语言的文本训练集训练得到分类器，并对另一种语言未标注类别的文本进行分类的过程。相比于传统的文本分类，跨语言文本分类是一个较新的领域，研究起步较晚。2003 年 Bel 等人[1]第一次正式提出跨语言文本分类这一学术性概念，并指出：跨语言文本分类是指在无须人工干预的情况下将现有的文本分类系统从单语言扩展到两种或多种语言。国内外研究者先后提出了基于双语词典、机器翻译、潜在主题模型等跨语言文本分类方法。Bel 等人[1]对每个类别源语言文档的前 n 个词构成类别的特征词，之后使用双语词典将未分类文档翻

收稿日期：2018-04-03；**修回日期：**2018-05-07　　**基金项目：**国家语委“十二五”科研规划项目（YB125-178）

作者简介：田明杰（1990-），男（朝鲜族），吉林延吉人，硕士研究生，主要研究方向为自然语言处理；崔荣一（1962-），男（通信作者），教授，博士，主要研究方向为智能计算、模式识别、机器学习、自然语言处理（cuirongyi@ybu.edu.cn）.

译成目标语言，最后通过相似度比较进行分类。Olsson 等人[2]通过概率的双语词典方式将英文训练文档翻译成捷克语文档。基于机器翻译的方法是将所有文本翻译为另一种语言后进行跨语言文本分类，包括源语言的训练集翻译为目标语言进行分类和目标语言的测试集翻译为源语言进行分类。Rigutini 等人[3]提出了一种结合机器翻译与 EM 算法的跨语言文本分类方法，对英语和意大利语文档进行了分类。Wei 等人[4]在跨语言情感分类中的结构对应学习方法中通过机器翻译的方式翻译中枢特征词。Mimno 等人[5]提出 PLTM 主题模型，对平行语料库与可比较语料库建模，进行了机器翻译任务与跨语言主题跟踪。Ni 等人[6]提出从维基百科（Wikipedia）中英可对比语料中挖掘多语言主题的方法，作者利用 LDA（latent Dirichlet allocation）主题模型为多语言主题建模，并把多语言文本投射到潜在主题空间里进行了跨语言文本分类。

基于词典的方法的最大缺陷在于词的歧义性和前 n 个特征词选择对翻译带来了困难，而且其策略是将领域内的关键词收录到词典中，当出现词典未收录的新词时，无法对领域内关键词进行收录。与借助双语词典的方法相比，采用机器翻译系统可以获取更多的语义信息，但是翻译系统的准确率会较大影响文档分类质量。基于潜在主题模型的跨语言文本分类方法中，每个潜在主题没有明确的定义，缺乏可解释性。

本文利用科技文献、新闻报道中的多标签信息与 LDA 主题模型在跨语言文本处理中的应用，提出带标签双语主题模型（labeled bilingual topic model, LBTM），旨在解决以往 LDA 主题模型中潜在主题概念的不明确性，使主题模型中的主题有更明确的语义和更好的可解释性。同时利用本文提出的主题模型在多个文档里挖掘出的共同“主题”，发现文档之间的关联性与相关性，最后通过这些“主题”对新的文档进行标签的推荐。

1　LDA 主题模型

LDA 主题模型[7]是由 Blei 于 2003 年提出的一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。Blei 认为一篇文章的每一个词的生成过程是：以一定概率选择某个主题，并从这个主题中以一定概率选择某个词语；文档到主题服从多项式分布，主题到词服从多项式分布。这种假设有利于大规模数据处理中的空间降维，即把文档投影到主题空间。LDA 主题模型在文档-主题与主题-词项的分布上引入了 Dirichlet 先验参数[8]，解决了在处理大规模语料库时出现的过拟合问题。

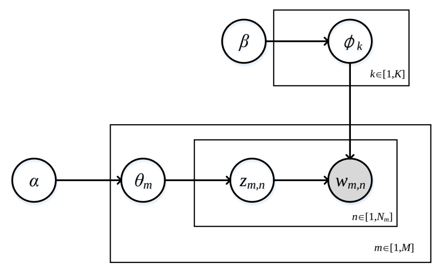


图 1　LDA 主题模型的概率图模型

LDA 主题模型是典型的有向概率图模型[9]，其概率图模型如图 1 所示，$w_{m,n}$ 表示第 m 篇文档第 n 个词，α 、β 为根据经验给定的 Dirichlet 分布的超参数，$z_{m,n}$ 是单词 $w_{m,n}$ 所对应的主题，参数 θ_m 为第 m 篇文档在主题上的分布，参数 ϕ_k 为第 k 个主题在词上的分布；给定文档集合 D，D 是一个 M 篇文档构成的集合，第 m 篇文档包含 N_m 个词，假设文档集 D 的主题数目为 K 个，LDA 主题模型生成文档的过程如下：

(1)　获取文档的长度 $N \sim Poission(\xi)$

(2)　选择文档在主题上的分布 $\theta \sim Dir(\alpha)$

(3)　对于每个在文档中的每个位置(n=1 to N)

　(3.1) 选择主题 $z_n \sim Multinomial(\theta)$

　(3.2) 以 $p(w_n \mid z_n, \beta)$ 的概率选择词 w_n

这里 $Poission(\cdot)$ 、$Dir(\cdot)$ 和 $Multinomial(\cdot)$ 分别表示泊松分布、Dirichlet 分布和多项式分布。

在构建 LDA 主题模型的过程中需要进行模型参数的估计，常用的方法主要有变分贝叶斯推理[10,11]、EM 算法[12]和 Collapsed Gibbs 采样[13]等方法。基于 Gibbs 采样的参数推理方法容易理解且实现简单，能够非常有效地从大规模文档集中采样主题，其参数估计过程可以被认为是文档生成的逆过程，即在已知文档集的情况下，通过估计得到参数值。根据图模型，可以得到一篇文档概率值为：

$$p(w \mid \alpha,\beta)=\int p(\theta \mid \alpha)(\prod_{n=1}^{N}\sum_{z_n} p(z_n \mid \theta)P(w_n \mid z_n,\beta))d\theta \tag{1}$$

可以通过积分避开实际待估计的参数，转而对每个词的主题进行采样。单词序列下主题序列的条件概率计算如下：

$$p(z_i=k \mid \overrightarrow{z_{\neg i}},\vec{w})=\frac{p(\vec{w},\vec{z})}{p(\vec{w},\overrightarrow{z_{\neg i}})}=\frac{n_{k,\neg i}^{t}+\beta_t}{\sum_{v=1}^{V}(n_k^v+\beta_v)-1}\frac{n_{m,\neg i}^{k}+\alpha_k}{\sum_{l=1}^{L}(n_m^l+\alpha_l)-1} \tag{2}$$

其中：z_i 表示第 i 个单词对应的主题；$\neg i$ 表示不包括其中的第 i 项；n_k^t 表示 k 主题中出现词 t 的次数；β_t 是词 t 的 Dirichlet 先验；n_m^l 表示文档 m 出现主题 l 的次数；α_l 是主题 l 的 Dirichlet 先验。

一旦获得每个词项的主题编号，参数可通过以下公式计算：

$$\phi_{k,t}=\frac{n_k^t+\beta_t}{\sum_{v=1}^{V}(n_k^v+\beta_v)} \tag{3}$$

$$\theta_{m,l}=\frac{n_m^k+\alpha_k}{\sum_{l=1}^{L}(n_m^l+\alpha_l)} \tag{4}$$

其中：$\phi_{k,t}$ 表示主题 k 中词 t 的概率；$\theta_{m,k}$ 表示文档 m 中主题 k 的概率。

2　带标签双语主题模型

LDA 主题模型将高维度的词项信息以低维的“潜在”主题形式来表征，可以捕捉文档的语义信息。但每个“潜在”主题没有明确含义，主题概念不明确，缺乏可解释性。本文利用科技文献、新闻报道中的多标签信息（比如论文中的关键词），对

LDA 主题模型进行改进，提出带标签双语主题模型（LBTM）。把科技文献与新闻报道中的标签视为主题，使主题有明确的含义和解释性，把“潜在”的主题实例化，赋予明确的内涵。使用本文提出的模型对文档进行建模，文档以“明确”的主题表示，对文档有更具体的说明；使用“明确”的主题表示文档集中的单词集合，对概括的单词集合有好的代表方式。通过带标签双语主题模型建模，文档集中每篇文档有着“明确”的主题的概率分布，可表示为向量空间模型中的向量，以实现跨语言文本分类与标签的推荐。

2.1　基本思想

假设文档集由 M 篇文档组成，每篇文档内容由两种语言 $L1$、$L2$ 描述，每种语言所讲述的内容是同样的。带标签双语主题模型使用一组与语言无关的“通用”主题，对文档的 2 种不同语言描述内容进行建模，每个“通用”主题都有 2 种不同的表示形式，每种表示形式与一种语言相对应。带标签双语主题模型的概率图模型如图 2 所示。

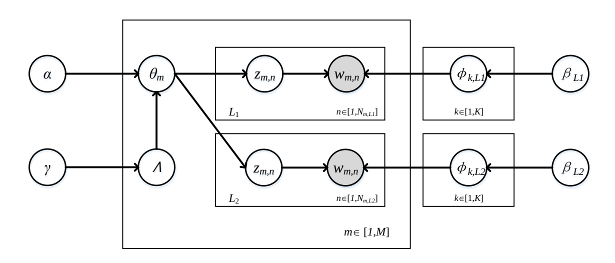


图 2　带标签双语主题模型概率图模型

其中 $w_{m,n}$ 表示第 m 篇文档第 n 个词，α 为根据经验给定的 Dirichlet 分布的超参数，β_{L_j} 为主题在语言 L_j（$j=1,2$）上的 Dirichlet 分布的超参数，γ 为文档-主题之间约束之间伯努利分布的超参数，Λ 表示文档与主题之间关系约束，且每篇文档与主题间的约束都是独有的，$z_{m,n}$ 是单词 $w_{m,n}$ 所对应的主题，参数 θ_m 为第 m 篇文档在主题上的分布，ϕ_{k,L_j} 为第 k 个“通用”主题在语言 L_j 的词项上的分布；给定文档集合 D，D 是有 M 篇文档构成的集合，每篇文档包含两种语言的内容表示，第 m 篇文档的语言 L_j 部分包含 $N_{m,Lj}$ 个单词，假设文档集 D 的主题数目为 K 个，带标签双语主题模型的生成文档的过程如下：

(1)　对于每个“通用”主题 z，（z=1，2，…，K）

(2)　对于每种语言 L_j（ j=1,2）

(3)　选择词项的分布　$\phi_{z,L_j} \sim Dir(\beta_{L_j})$

(4)　然后对文档集里的每篇文档 m

(4.1) 对于每个“通用”主题 z，（z=1，2，…，K）

(4.1.1) 选择 $\Lambda_{m,k} \in \{0,1\} \sim bernoulli(\gamma)$

(5)　选择在“通用”主题上的分布　$\theta \sim Dir(\alpha) \mid \Lambda$

(5.1) 对文档里的第 n 个词

(5.1.1)　选择主题号 $z_{m,n} \sim Multinomial(\theta_m)$

(5.1.2)　选择词 $w_{m,n,L_j} \sim Multinomial(\phi_{m,n,L_j})$

2.2　参数估计

在参数估计阶段，针对提出的主题模型标签与双语语言特性，对 Gibbs 采样方法进行修改。单词序列下的主题序列条件概率从单语的 $p(z_i = k \mid \vec{z}_{\neg i}, \vec{w})$ 扩展为双语的 $p(z_{i,L_j} = k \mid \vec{z}_{\neg i,L_j}, \vec{w}_{L_j})$，其中 $\vec{w}_{L_j}$ 表示语言 L_j 中所有词项组成的向量；$\vec{z}$ 表示这些词项的主题分配；$\vec{z}_{\neg i,L_j}$ 表示在不考虑当前 L_j 语言中第 i 个词的主题分配。条件概率的公式如下：

$$p(z_{i,L_j} = k \mid \vec{z}_{\neg i,L_j}, \vec{w}_{L_j}) = \frac{n_{k,\neg i,L_j}^{t} + \beta_{L_j}^{t}}{\sum_{v=1}^{V_{L_j}} (n_{k,L_j}^{v} + \beta_{L_j}^{v}) - 1} \frac{\sum_{j=1}^{2} n_{m,\neg i,L_j}^{k} + \alpha_k}{\sum_{l=1}^{L} (\sum_{j=1}^{2} n_{m,L_j}^{l} + \alpha_l) - 1} \tag{5}$$

其中：$n_{k,\neg i,L_j}^{t}$ 表示不考虑当前词项 t 的当前主题分配的情况下，语言 L_j 中词项 t 的主题分配到 k 的次数；$\sum_{v=1}^{V_{L_j}} n_{k,L_j}^{v} - 1$ 表示不考虑当前词项 t 的当前主题分配的情况下，语言 L_j 中所有词项的主题分配到 k 的次数；V_{L_j} 表示语言 L_j 的词典；$n_{m,\neg i,L_j}^{k}$ 表示不考虑当前词项 t 的当前主题分配的情况下，文档 m 中语言 L_j 所有词项的主题分配到 k 的次数；$\sum_{l=1}^{L}\sum_{j=1}^{2} n_{m,L_j}^{l} - 1$ 表示忽略当前词项 t 的情况下，文档 m 中两种语言中单词的总个数。获得文档每种语言每个单词的主题分配以后，带标签双语主题的文档在主题上的表示如下：

$$\theta_{m,k,L_j} = \frac{n_{m,L_j}^{k} + \alpha_k}{\sum_{l=1}^{L} (n_{m,L_j}^{l} + \alpha_l)} \tag{6}$$

2.3　估计新文档的主题分布

对于一篇新的文档，它在主题上的分布可以通过已经训练完成的模型参数来预测，并将新文档投射到“主题”维度上的分布。此时我们需要计算条件概率 $p(z_{i,L_j}^{d} = k \mid \vec{z}_{\neg i,L_j}^{d}, \vec{w}_{\neg i,L_j}^{d}, \vec{z}_{L_j}, \vec{w}_{L_j})$，

其中 $\vec{w}_{L_j}^{d}$ 表示文档 d 在 Bag-of-Word 模型下的向量；$\vec{w}_{L_j}^{d}$ 中当前词项 t 的主题分配依赖于文档中其他词项的当前主题分配与所有词项的主题分配，计算公式如下：

$$\begin{aligned} & p(z_{i,L_j}^{d} = k \mid \vec{z}_{\neg i,L_j}^{d}, \vec{w}_{\neg i,L_j}^{d}, \vec{z}_{L_j}, \vec{w}_{L_j}) \\ & = \frac{n_{k,L_j}^{t} + \beta_{L_j}^{t} + n_{\neg i,k,L_j}^{d,t}}{\sum_{v=1}^{V_{L_j}} (n_{k,L_j}^{v} + \beta_{L_j}^{v} + n_{k,L_j}^{d,v}) - 1} \frac{n_{\neg i}^{d,k} + \alpha_k}{\sum_{l=1}^{L} (n^{d,l} + \alpha_l) - 1} \end{aligned} \tag{7}$$

其中：$n_{\neg i,k,L_j}^{d,t}$ 表示不考虑当前词项 t 在文档 d 的第 i 个位置的主题分配的情况下，词项 t 的主题分配 k 的次数；$n_{\neg i}^{d,k}$ 表示不考虑当前词项 t 在文档 d 的第 i 个位置的主题分配的情况下，文档 d 中其他位置的词项的主题分配到 k 的次数。最后可以计算出新的文档在主题上的分布 $\vec{\theta}^{d} = \{\theta_1^d, \theta_2^d, \ldots, \theta_K^d\}$，其中每个分量的表示如下：

$$\theta_k^d = \frac{n^{d,k} + \alpha_k}{\sum_{l=1}^{L} (n^{d,l} + \alpha_l)} \tag{8}$$

2.4　与“潜在”主题模型的区别

相比于 LDA 主题模型等“潜在”主题模型，本文提出的带标

签双语主题模型利用与文档链接的多标签数据，实例化“潜在”主题，使主题的意义不再是“隐含”的，而是“明确”的。与传统“潜在”主题模型区别如下：

a)主题数 *K*。主题数 *K* 的大小的确定是 LDA 主题模型的难点之一，需要从一定范围内通过实验结果选择 K 的值。而带标签双语主题模型中主题数 *K* 是确定的，即语料库中所有文档的标签中去重后唯一标签数；

b)文档的向量表示。在确定主题数 *K* 的前提下，“潜在”主题模型中每篇文档每个单词的主题采样范围为 1~*K*，即所有主题，文档是基于文档内单词与主题的所属关系表示的，所以文档的向量表示中每个主题分量的值都可能不为 0。带标签双语主题模型中，每个文档与该文档标注的标签有约束关系，文档向量（文档到主题的分布）的表示 θ 中，只与该文档有约束关系的主题分量值不为 0，其余为 0；

c)文档生成过程。在文档到主题的分布 θ 确定阶段，传统“潜在”主题模型,根据 Dirichlet 先验 α 选取文档中的主题分布 θ。带标签双语主题模型中，每个文档与主题（标签）间要么有约束、要么没有约束，所以根据伯努利先验 γ 来确定文档到主题的分布；

d)采样范围。在每次迭代中需要计算每个文档中每个单词与所有主题间的条件概率 $p(z_i=k\,|\,\overrightarrow{z_{\neg i}},\vec{w})$。传统的“潜在”主题模型，由于文档与主题之间没有约束关系，每篇文档中的每个单词主题采样范围为 *K*。带标签双语主题模型中，每篇文档与标签之间有固定的归属约束 Λ，每篇文档中每个单词主题采样范围为与该文档有约束关系的主题（标签）集合；

e)采样计算复杂度。由于采样范围的不同，在每一次的迭代更新单词的主题分配的过程中，传统“潜在”主题模型需要计算每篇文档中每个单词与所有主题间的条件概率。带标签双语主题模型中，只需要计算与该文档有关联的主题间的条件概率。本文提出的模型在采样过程中的计算效率上有一定的优势；

f)新文档的主题分布的推断。对于一篇新的文档，通过已经训练完成的模型参数推断在主题上的分布，推断过程中通过计算新文档中每个单词与主题间的条件概率 $p(z_{i,L_j}^d=k\,|\,\vec{z}_{\neg i,L_j}^{d},\vec{w}_{\neg i,L_j}^{d},\vec{z}_{L_j},\vec{w}_{L_j})$ 来对分配的主题进行采样。传统“潜在”主题模型采样的主题范围为 *K*。带标签双语主题模型中，采样范围为训练阶段当前单词所分配到的所有主题。

本文提出的带标签双语主题模型中，每个主题有着“明确”含义。其“明确性”确定了主题数与每篇文档中单词的主题采样范围，并且在新文档主题推断阶段减少了与单词所属主题采样范围。

3 实验结果及分析

3.1 跨语言文本分类

为了验证本文提出的带标签双语主题模型在跨语言文本分类任务上的有效性和可行性，用训练集的汉语与朝鲜语科技文献训练出分类器，并使用分类器对测试集的汉语与朝鲜语的文档分别进行跨语言文本分类。语料库是平行语料库，所以汉语与朝鲜语文档的内容是相同的，即在语义上是对齐的。作为对比，与不使用标签信息的传统“潜在”LDA 主题模型——文献[6]的方法进行实验对比。

3.1.1 数据集

本实验所使用的双语语料资源为延边朝鲜族自治州科技局中朝科技文献平行语料库，科技文献包含汉语与朝鲜语句子级别对齐的 9 000 篇论文的关键词与摘要部分，对齐语料如图 3 所示。其中生态类 6 000 篇、航空航天类 3 000 篇，类别（生态与航空航天）的定义标准为收录论文的期刊类别，其中各类别选取训练集与测试集的比例为 9:1。汉语摘要部分使用 ICTLAS 分词系统进行分词，朝鲜语摘要部分使用 Hannanum 分词系统进行分词。

基于改进速度增益的中段变轨制导方法研究

关键词：改进速度增益; 虚拟目标; 中段变轨制导; 需要速度; 状态预估

摘要：变轨技术的核心是控制单元的算法部分。基于对传统速度增益制导律的认识,提出了一种基于改进的速度增益制导方法,解决中段改变打击目标的实时控制问题。通过建模、仿真,验证了算法的可行性。

속도 증가 개선 기반의 중간 궤도 변화시 유도법칙 연구

키워드:속도 증가 개선: 가상표적: 중간 궤도 변화의 유도: 속도 수요: 상태 예측

요약:궤도 변화 기술의 핵심은 제어장치의 알고리즘 부분에 있다.기존에알고 있는 속도 증가의 유도 법칙에 대한 인식을 기반으로 한가지의 개선한 속도 증가 유도 법칙을 제안하였으며 중간 궤도에서의 공격 대상을 변경할때실시간 제어 문제점을 해결하였다.또한, 구축 모델과 시뮬레이션을 통하여 알고리즘의 실행 가능성을 검증하였다.

图 3　中朝科技文献平行语料

3.1.2 评价指标

对分类结果的评测，本文采用 *Macro-F1* 值与 *Micro-F1* 值两个指标对分类性能进行评价。*Macro-F1* 实现对每一个类统计指标值，然后再对所有内求算术平均值。*Micro-F1* 是对测试集中的每一个实例不分类别进行统计，建立全局的混淆矩阵，然后计算相应指标。具体定义如下:

$$Macro_F1=\frac{2*Macro_precision*Macro_recall}{Macro_precision+Macro_recall} \tag{9}$$

$$Micro_F1=\frac{2*Micro_precision*Micro_recall}{Micro_precision+Micro_recall} \tag{10}$$

其中 *precision* 为精确率，指的是被正确分类的文档与所有被分类到相应类别的文档的比值，*recall* 为召回率，指的是被正确分类的文档与实际属于相应类别的文档的比值。具体定义如下:

$$Macro_precision=\frac{1}{|C|}\sum_{i=1}^{|C|}precision_i \tag{11}$$

$$Macro_recall=\frac{1}{|C|}\sum_{i=1}^{|C|}recall_i \tag{12}$$

$$Micro_precision=\frac{\sum_{i=1}^{|C|}TP_i}{\sum_{i=1}^{|C|}TP_i+FP_i} \tag{13}$$

$$Micro_recall=\frac{\sum_{i=1}^{|C|}TP_i}{\sum_{i=1}^{|C|}TP_i+FN_i} \tag{14}$$

其中：*C* 代表训练集中所有类别的集合，|*C*|代表类别的数量，*TP* 代表被正确分类到第 *i* 个类别的文档数量，*FP* 代表被错误

的分类到第 i 个类别的文档数量，*FN* 代表实际属于第 i 个类别，但是被错误分类到其他类别的文档数量。

Macro-F1 是每一个类别性能指标的算术平均，*Macro-F1* 值的结果极易受到小样本类别的影响。*Micro-F1* 是各个文档性能指标的算术平均，*Micro-F1* 值的结果容易受到文档集中文档数较多的类别影响[14,15]。*Micro-F1* 值与 *Macro-F1* 值作为两个可以综合度量文本分类性能的评价指标，在文本分类的研究中被广泛应用。本文在验证提出方法的分类性能时也将采用这两个评价标准。

3.1.3 参数设定

主题模型需要对模型中的参数进行预先的设定。需要确定主题数 K、Dirichlet 先验参数 α 与 β，训练与测试迭代次数。带标签双语主题模型与“潜在”LDA 主题模型参数设置如表 1 所示。

表 1　对比实验参数设定

	带标签双语主题模型	“潜在”LDA 主题模型
主题数 K 总计	18170	400
先验参数 α	50/K	50/K
先验参数 β	0.01	0.01
训练迭代次数	1000	1000
测试迭代次数	100	100

其中，本文提出的带标签双语主题模型的主题数 K 为确定的，主题数为 18170，即训练集中去重后的唯一标签数。“潜在”LDA 主题模型的主题数 K 设置为 400，进行对比试验。在训练阶段与测试阶段，分别进行 1000 次与 100 次的迭代，以期文档在主题上的分布与主题在词项上的分布变化达到收敛状态。

3.1.4 实验结果及分析

为了比较分类精度，在中朝科技文献语料上分别使用本文提出的带标签双语主题模型与传统“潜在”主题模型文献[6]的方法进行比较。通过主题模型对文档集的建模，训练集与测试集中的文本都被表示为主题上的分布。本实验使用朴素贝叶斯分类器对双语文档进行跨语言文本分类，具体的分类任务包括：通过朝鲜语训练文档训练的分类器，对汉语测试文档进行分类；通过汉语训练文档训练的分类器，对朝鲜语文档进行分类。训练集与测试集以 9:1 的比例随机抽取，统计训练集中的标签与词项，构建标签索引与词典。利用训练集的标签（关键词）与内容部分（摘要），对带标签双语主题模型进行参数估计。利用带标签双语主题模型对测试集文档进行推断，得到“通用”主题上的分布 $\vec{\theta}^{d}$。利用训练集数据训练朴素贝叶斯分类器，最终对测试集文档进行分类。

本文提出的带标签双语主题模型与“潜在”LDA 模型实验所得的分类精度与训练消耗时间如表 2 所示。

表 2　对比实验结果

评价指标 / 模型	*Micro-F1* KOR->CHN	*Micro-F1* CHN->KOR	*Macro-F1* KOR->CHN	*Macro-F1* CHN->KOR	时间消耗
本文	94.79%	94.81%	92.31%	92.41%	8hours
文献[6]方法	92.32%	92.76%	94.04%	94.37%	48hours

表中 KOR->CHN 表示用科技文献中朝鲜语部分的关键词与摘要训练出的分类器对汉语部分进行分类的结果（其余表示法以此类推）。

本文提出的带标签双语主题模型在跨语言文本分类任务中 *Micro-F1* 值最高达到 94.81%，*Macro-F1* 值最高达到 92.41%，可应用于实际的科技文献的跨语言自动文本分类工作中。与“潜在”主题模型的对比中,本文提出的模型 *Micro-F1* 值高于对比实验，*Macro-F1* 值低于对比实验。在 6000 篇的生态类科技文献里，本文提出的模型分类精确度高于对比实验，而 3000 篇的航空航天类科技文献中，对比实验分类精确率高于文本提出的模型。正好验证了，*Micro-F1* 值容易受到文档集中文档数较多的类别影响，*Macro-F1* 值容易受到小样本类别的影响的特点。对模型评价，需要综合两种指标。

跨语言文本分类过程中，首先对训练集进行建模，每一次训练迭代中根据式（5）计算每篇文档里每个单词分配到每个主题的概率，用 Gibbs 采样方法采样更新分配的主题，训练迭代完成后根据式（6）获得用于跨语言文本分类的主题模型的参数 θ（训练文档在主题下的分布）；其次对于测试集中的一篇新的文档，根据上一步骤训练获得的模型参数 θ，结合新文档中出现的词项，初始化新文档的主题分布，并在每一次测试迭代中根据式（7）计算新文档里每个单词分配到每个主题的概率，与训练过程一样用 Gibbs 采样方法采样更新分配的主题，最后根据式（8）获得新文档在主题下的分布。训练与测试过程中，迭代次数的设定参考表 1。训练与测试阶段总的时间消耗上，本文提出的带标签双语主题模型总共耗时 8 h，“潜在”主题模型共耗时 48 h，总的时间消耗比为 1:6。在训练阶段，本文提出的模型每篇文档中每个单词的主题采样范围给定，一般为每篇论文的关键词个数（5~6 个），而对比实验中，文档与主题没有约束关系，需要对所有的主题计算条件概率概率并采样，增加了计算复杂度；在新文档的主题推断阶段，需要计算新文档每个单词与主题间的条件概率，本文提出的模型只需要计算训练阶段当前单词所分配到的所有主题间的概率即可，“潜在”主题模型需要计算与所有主题间的条件概率，增加了计算复杂度。

综上，本文提出的方法在得到较高的分类精确度的同时相比于对比方法节省了大量时间。

3.2　标签推荐

对“潜在”主题的实例化与明确化，带标签双语主题模型可以应用于标签的推荐。通过本文提出的模型对没有标注标签的新文档进行主题的推断，得到文档在主题上的分布。把文档表示成以主题为维度的向量时，每个分量值的意义为：文档中所有单词里，属于该主题的单词的占比。分量值越大说明文档与该主题的相关性越大。

本实验将使用与跨语言文本分类任务相同的文档集。具体

实现方法是：对测试集文档完成推断，最终表示为主题上的分布，抽取分量值最大的前两个主题作为推荐的标签，与原有的标签进行比对。中朝科技文献中题目、关键词与本文提出的模型推荐的标签结果如表 3、4 所示。

表 3 中中文文献“膜下滴灌技术生态-经济与可持续性分析—以新疆玛纳斯河流域棉花为例”中，论文的关键词与模型推荐的标签里都含有“棉花”。表 4 中朝鲜语文献“LFM 광대역 레이더 신호의 블라인드 압축 센싱 모델”(LFM 宽带雷达信号的盲压缩感知模型)中，论文的关键词与模型推荐的标签里都含有“압축 센싱”(压缩感知)。并且在其他文献的关键词语模型推荐标签之间，普遍存在着语义上的关联。

表 3　中文科技文献标签推荐结果

论文标题	论文关键词	模型推荐的标签
膜下滴灌技术生态-经济与可持续性分析—以新疆玛纳斯河流域棉花为例	棉花；可持续分析；Bossel 理论；	数值模拟；棉花；
施肥对板栗林地土壤 N2O 通量动态变化的影响	施肥；水溶性有机碳；	土壤有机碳；生长；
马赫数 4 下氢气自燃辅助乙烯点火实验研究	直连式脉冲燃烧风洞；点火试验；	点火；亚燃模态；

表 4　朝鲜语科技文献标签推荐结果

论文标题	论文关键词	模型推荐的标签
우주 테더 로봇의 표적 접근 과정에 대한 통합 자세제어 (空间绳系机器人逼近过程的位姿一体化控制)	우주 테더 (空间绳系)；통합 자세제어 (位姿一体化控制)；최적 제어 （最优控制)；	우주 테더 로봇 (空间绳网机器人)；자태 제어 (姿态控制)；
LFM 광대역 레이더 신호의 블라인드 압축 센싱 모델 (LFM 宽带雷达信号的盲压缩感知模型)	압축 센싱 （压缩感知）；선형 주파수 변조 신호 （线性调频信号）；분수계 푸리에 변환 （分数阶傅里叶变换）；	신호정렬 （信号分选）；압축 센싱 (压缩感知)；
티베트 써지라 Abies georgei var. smithii forest 에 토양 선충 집단의 특징 （西藏色季拉山急尖长苞冷杉林土壤线虫群落特征）	선충 （线虫）；군집구조 （群落结构）；생물다양성 （生物多样性）；티베트 （西藏)；	군집구조 （群落结构）；종 다양성 （物种多样性）；

每篇论文的关键词由论文作者人工添加，对事物的看法不可避免地会有差别，所以推荐的结果无法与论文作者添加的关键词精确对齐，因而利用语义上的关联，可以进行辅助性标签推荐。

4　结束语

本文利用科技文献、新闻报道中的多标签信息，结合 LDA 主题模型，提出了带标签双语主题模型。本文提出的带标签双语主题模型有如下的特点：

a）相比于传统“潜在”LDA 主题模型，对主题进行“实例化”，有了更明确的内涵和更好的解释性；

b)在训练参数与新文档主题分布推断阶段，由于采样范围给定，在参数估计与文档推断速率方面优于传统“潜在”LDA 主题模型；

c)以中朝科技文献平行语料库为语料，跨语言文本分类任务中，*Macro-F1* 值达到 92.41%，*Micro-F1* 值更是达到了 94.81%，可适用于实际应用；

d)依据每个主题的明确的含义，可用于辅助性的标签推荐。

在此模型的基础上，下一步研究工作如下：

a)通过提取各标注类别的领域特征词，提高跨语言文本分类精度；

b)受限于语料库的语种，现阶段只能应用于双语语料的主题建模。如果有涵盖更多语言的带多标签的平行语料库，则将模型扩展到多种语言建模的主题模型。

参考文献：

[1] Bel N, Koster C H A, Villegas M. Cross-lingual text categorization [C]// Proc of the 7th European Conference on Research and Advanced Technology for Digital Libraries. Berlin: Springer, 2003: 126-139.

[2] Olsson J S, Oard D W, Hajic J. Cross-language text classification [C]// Proc of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2005: 645-646.

[3] Rigutini L, Maggini M, Liu Bing. An EM based training algorithm for cross-language text categorization [C]// Proc of IEEE//WIC//ACM International Conference on Web Intelligence. Washington DC: IEEE

Computer Society, 2005: 529-535.

[4] Wei Bin, Pal C. Cross lingual adaptation: an experiment on sentiment classifications [C]// Proc of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2010: 258-262.

[5] Mimno D, Wallach H M, Naradowsky J, *et al*. Polylingual topic models [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2009: 880-889.

[6] Ni Xiaochuan, Sun Jiantao, Hu Jian, *et al*. Mining multilingual topics from Wikipedia [C]// Proc of the 18th International Conference on World Wide Web. New York: ACM Press, 2009: 1155-1156.

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3 (1): 993-1022.

[8] 徐谦，周俊生，陈家骏. Dirichlet过程及其在自然语言处理中的应用 [J]. 中文信息学报, 2009, 23 (5): 25-32//46. (Xu Qian, Zhou Junsheng, Chen Jiajun. Dirichlet process and its applications in natural language processing [J]. Journal of Chinese Information Processing, 2009, 23 (5): 25-32//46.)

[9] 徐戈，王厚峰. 自然语言处理中主题模型的发展 [J]. 计算机学报, 2011, 34 (8): 1423-1436. (Xu Ge, Wang Houfeng. The Development of Topic Models in Natural Language Processing [J]. Chinese Journal of Computers, 2011, 34 (8): 1423-1436.)

[10] Beal M J. Variational algorithms for approximate Bayesian inference [D]. London: University of London, 2003.

[11] Fang Anjie, Macdonald C, Ounis I, *et al*. Exploring time-sensitive variational Bayesian inference LDA for social media data [C]// Proc of the 39th European Conference on Information Retrieval. Berlin: Springer, 2017: 252-265.

[12] 王爱平，张功营，刘方. EM 算法研究与应用 [J]. 计算机技术与发展, 2009, 19 (9): 108-110. (Wang Aiping, Zhang Gongying, Liu Fang. Research and application of EM algorithm [J]. Computer Technology and Development, 2009, 19 (9): 108-110.)

[13] Yerebakan H Z, Dundar M. Partially collapsed parallel Gibbs sampler for Dirichlet process mixture models [J]. Pattern Recognition Letters, 2017, 90: 22-27.

[14] 张启蕊，张凌，董守斌，等. 训练集类别分布对文本分类的影响 [J]. 清华大学学报：自然科学版, 2005, 45 (S1): 1802-1805. (Zhang Qirui, Zhang Ling, Dong Shoubin, *et al*. Effects of category distribution in a training set on text categorization [J]. Journal of Tsinghua University: Science and Technology, 2005, 45 (S1): 1802-1805.)

[15] Luo Le, Li Li. Defining and evaluating classification algorithm for high-dimensional data based on latent topics [J/OL]. PLoS One, 2014, 9 (1): e82119. (2014-01-09) [2018-05-05]. https://doi. org/10. 1371/journal. pone. 0082119.